网络招聘文本技能信息自动抽取研究*

俞琰[1,2]　陈磊[1]　姜金德[3]　赵乃瑄[1]

[1] 南京工业大学　南京 210009　[2] 东南大学成贤学院计算机工程系　南京 211816

[3] 南京晓庄学院商学院　南京 211171

摘要：［目的/意义］针对目前网络招聘文本手工抽取技能信息无法满足大数据量分析要求的问题，提出一种针对大量网络招聘文本的技能信息自动抽取方法。［方法/过程］根据网络招聘文本的特点，利用依存句法分析选取候选技能，然后提出领域相关性指标衡量候选技能，将其融入传统的术语抽取方法之中，形成一种网络招聘文本技能信息自动抽取方法。［结果/结论］实验表明，本文提出的方法能够从网络招聘文本中自动、快速、准确地抽取技能信息。

关键词：网络招聘文本　技能信息自动抽取　术语抽取

分类号：G202

DOI：10.13266/j.issn.0252－3116.2019.13.011

1　引言

随着互联网的普及，网络招聘成为企业招聘的主流方式。网络招聘文本中常包含招聘岗位所需技能的具体描述。通过网络招聘文本技能信息抽取与分析，可以了解当前就业市场对某个领域人才技能的需求，为高校制定符合企业需求的专业人才培养方案提供决策支持，在一定程度上可以化解大学生找工作难、企业招人难的问题。特别地，从非结构化网络招聘文本中抽取结构化技能信息是进行网络招聘文本技能信息分析的基础。图 1 为网络招聘文本所包含的技能信息抽取示例，其中“关系型数据库”“SQL 优化”“HTTP 协议”等为岗位所要求的技能信息。然而，目前相关研究通常采用手工方法从网络招聘文本中抽取技能信息[1-7]。显然，手工方法很难满足高速发展的信息时代下大数据量网络招聘文本技能信息分析的要求。

技能信息为特定领域中特定岗位对所需人才的专业知识和技术的要求，其本质为描述特定领域中知识活动理论概念的术语。因此，网络招聘文本技能信息抽取任务可以借鉴术语抽取研究中的方法。特别地，C-value 方法[8]是一种常见的、简单高效的术语抽取方法。然而，由于 C-value 方法主要基于词串在语料集中

......<em class="i4">04-21 发布<div class="clear"></div><div class="clear"></div></div></div></div><div class="tBorderTop_box"><h2>职位信息</h2><div class="bmsg job_msg inbox"><p>任职要求：</p><p>1、熟悉关系型数据库，并有一定的 SQL 优化经验 2、熟悉 HTTP 协议；3、具有良好的团队合作意识。</p><p>五险一金：享受齐全的社会保险，包括养老、医疗、失业、工伤、生育、以及住房公积金......

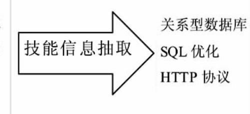


图 1　网络招聘文本技能信息抽取示例

出现的频次，无法有效地过滤网络招聘文本中出现频次高的非技能词串，如“开发 经验”“相关 专业”“熟悉 Linux”等。本文根据网络招聘文本的特点，提出首先利用依存句法分析选取候选技能，然后提出领域相关性概念，以度量候选技能的领域相关性，最后将其融入 C-value 方法之中，形成一种改进的 C-value 网络招聘文本技能信息自动抽取方法。实验表明，本文提出的方法能够从网络招聘文本中自动、快速、准确地抽取技能信息。

2　相关研究

2.1　网络招聘文本技能信息抽取

通常研究采用手工抽取技能信息的方法。如：I.

* 本文系教育部人文社科规划项目项目“大数据时代技能知识图谱构建研究”（项目编号：16YJAZH073）和国家社会科学基金一般规划项目“大数据时代支持创新设计的多维度多层次专利文本挖掘研究”（项目编号：17BTQ059）研究成果之一。

作者简介：俞琰（ORCID：0000－0002－9654－8614），教授，博士，E-mail：yuyanyuyan2004@126.com；陈磊，硕士研究生；姜金德（ORCID：0000－0002－5504－7493），教授，博士；赵乃瑄（ORCID：0000－0001－9072－7315），教授，博士。

收稿日期：2018－07－17　修回日期：2019－03－19　本文起止页码：105－113　本文责任编辑：王传清

Wowczko[1]手工抽取和映射招聘中的技能。J. Y. Kim等[2]手工分析数据科学家招聘信息,总结企业对数学职业家一职的专业以及学历要求。A. D. Mauro等[3]结合专家判断,分析2 700条大数据相关岗位信息,划分出4个领域相关的工作类型,并对每一个工作类型所需的技能和熟练程度要求进行评估。吕斌等[4]、李国秋等[5]手工调研300个情报职业招聘网页,分析社会组织的情报职业需求,以及社会组织中情报职业类型、职责和作用等。夏火松和潘筱听[6]对比硕博士论文以及招聘网站硕博士相关招聘信息,分析我国大数据在学界和业界的现状,发现我国大数据企业人才需求与高校学术研究之间的关系。黄崑等[7]手工抽取职位基本信息、岗位职责和任职要求,分析大数据岗位对人才知识和能力的要求,并对图书馆情报学科人才适应国内大数据环境下的培养方案提出建议。

显然,手工方法很难胜任大数据量、非结构化环境下的网络招聘信息分析要求。一些研究尝试使用基于外部资源、基于规则和基于统计的方法自动抽取网络招聘文本的技能信息。

基于外部资源的方法利用技能词典、维基百科等资源构建技能词典,然后与网络招聘文本的信息匹配抽取技能信息。如:M. S. Sodhi和B. G. Son[9]通过构建运筹学专业核心技能词典研究该专业招聘文本信息,以研究不同行业对运筹专业技能需求的差异。M. Zhao等[10]使用常规短语、领域专家预定义的各种术语分析招聘网页,使用维基进行去重和规范化。T. Xu等[11]从CSDN网站下载技能种类和具体技能,包括54个技能种类和1 729个技能,构建了技能字典。詹川[12]参考已有的电子商务专业术语,构建该专业的术语词典,从招聘文本中抽取高于一定频数的技能关键词并归类,分析电商各岗位的需求、技能整体需求和各个岗位特别需求的技能。夏立新等[13]利用中华教育在线职业大全、招聘网岗位分类、论文关键词构建专业、岗位和知识点词典,通过挖掘招聘文本信息,形成网络文本挖掘的“专业-岗位-知识点”的就业需求关系。

基于规则的方法利用技能信息出现位置的特征,人工构造规则模板,通过规则匹配实现技能信息抽取。如,M. Bastian等[14]利用逗号进行匹配,抽取LinkedIn网络招聘文本中的技能信息,将频次低于阈值的过滤,并使用维基进行技能规范化处理。

基于统计的方法主要利用语料库训练某个词作为技能信息的概率,若大于某一阈值,则认为是技能信息。如刘睿伦等[15]采用人工标注和改进的词频统计信息识别招聘信息文本中信息,利用聚类算法和轮廓稀疏确定实体转化成向量的最佳维度大小和聚类个数,抽取招聘网站关于大数据的工作岗位信息。

总体来说,目前网络招聘文本技能抽取仍然以手工抽取方法为主,不能适应大数据时代数据快速变化、数据量大的要求;而基于外部资源的方法存在外部资源更新较慢、覆盖面较窄的问题;基于规则和统计的方法则存在方法过于简单、结果不尽理想等问题。

2.2 术语抽取

术语抽取指从文本中自动发现术语的过程。目前,术语抽取方法可分为无监督方法和有监督方法两大类。无监督方法通常利用语言学与统计学相结合的方法,具有较少人工干预、较强的适用性和一致性等优点;有监督方法采用机器学习方法,通过学习训练文本特征,构造模型抽取术语,能够弥补无监督方法无法识别低频术语的缺陷,具有较高术语抽取准确率和召回率,但需要大规模人工标注语料作为训练数据,并且方法还不成熟,需要更多的尝试与验证[16]。目前网络招聘技能信息抽取任务没有大规模标注语料库,因此,本文着重研究使用无监督方法。

无监督方法通常首先从语料库中选取候选术语,然后利用统计信息计算候选术语成为术语的可能性。一般使用术语性和单元性度量候选术语成为术语的可能性。术语性衡量一个候选术语对领域知识的表达能力。单元性度量候选术语结构的稳定程度。特别地,C-value方法[8]是一种简单高效的基于术语性的术语抽取方法[17]。国内外有较多该方法的应用[18-19]。然而,由于C-value方法主要基于词串在语料集中出现的频次,无法有效过滤语料库中出现频次高的非术语词串。针对这个问题,较为典型的方法是引入互信息和邻接熵两种统计量,重构C-value目标函数[20]。互信息计算候选术语中各词依赖程度,互信息值越大,表明候选术语中各词的依赖程度越大,越可能是是术语[20]。邻接熵衡量候选术语左右邻接词的不确定性,其不确定性越大,表明其邻接词包含的信息越多,越可能是术语[20]。然而,网络招聘文本中一些非技能词串频繁共同出现,具有较高的互信息值,如“相关 专业”“工作经验”等,因此,互信息不能很好地衡量候选术语。同样地,网络招聘文本中一些高频非技能词串具有较高的邻接熵,如“熟练 使用”“具有 良好”等,也不能很好地衡量候选术语。

总体来说,术语抽取研究已经取得一定的成果,但

是如果直接将这些方法应用到网络招聘文本技能信息抽取之中,将造成技能信息抽取准确率和召回率低的问题。

3 网络招聘文本技能信息自动抽取方法

本文根据网络招聘文本的特点,首先利用依存句法分析选取候选技能,然后提出技能领域相关性概念,在欲抽取的目标领域网络招聘文本集的基础上,引入非目标领域网络招聘文本集,以度量技能信息的领域相关性,以改进 C-value 方法。方法流程如图 2 所示,主要包括预处理(第 3.1 节)、基于依存句法分析的候选技能选取(第 3.2 节)、C-value 值计算(第 3.3 节)、领域相关性度量(第 3.4 节)和融入领域相关性的 C-value 值计算(第 3.5 节)等步骤。

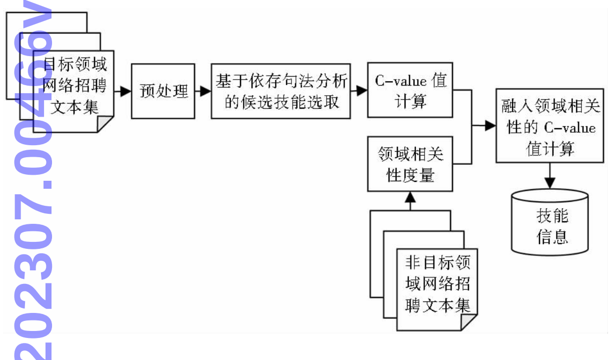


图 2　网络招聘文本技能信息自动抽取方法流程

3.1 预处理

由于招聘文本是非结构化的网页结构,而且除了包含技能等所需信息之外,还包括其他大量噪音信息,如广告、图片动画、与主题无关的超链接、脚本语言以及各类标签。因此,首先针对网页文本结构,使用 Beautiful Soup 等网页文本分析工具定位、解析网络内容,获得技能要求文本。然后,对获取的相关文本内容进行去重、英文大小写转化、去除特殊字符等操作。图 3 为网络招聘文本预处理示例。

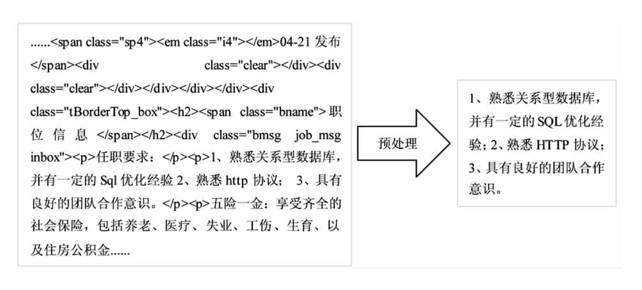


图 3　预处理示例

3.2 基于依存句法分析的候选技能选取

已有的研究通常利用连续名词、动词等词串选取候选技能[21]。然而,这种方法会包括大量噪声动词的非技能词串,如"熟练 使用""熟悉 HTTP 协议"等,从而造成最终技能抽取准确率低;但是如果选取不包含动词的候选技能词串,则可能遗漏一些候选技能,如"SQL 优化"中的"优化"一词为动词,从而造成最后技能抽取召回率低。

通过分析网络招聘文本可以发现,包含技能信息的文本通常为动宾结构,如"熟悉关系型数据库"。因此,本文提出利用依存句法分析,以剔除"熟悉"等噪声动词。依存句法分析通过语句单位内词语间的依存关系揭示词语间的语义修饰关系。其中,依存关系使用有向弧表示,由支配词指向其从属词,并且依存句法分析认为语句中的支配者是核心动词。根据依存语法公理[22],在一个语句中,依存句法分析将语句的线性结构层次化,构造成为依存树。图 4 为使用哈尔滨工业大学语言技术平台发布的依存句法分析器[23－24],分别对语句"熟悉关系型数据库""并有一定的 SQL 优化经验""熟悉 HTTP 协议"和"具有良好的团队合作意识"进行依存句法分析之后得到的依存树 T1、T2、T3 和 T4。图 4 中,Root 分别指向语句的核心动词"熟悉""有""具有",结点下的字母表示词性,v 表示动词、n 表示名词、c 表示连词、b 表示区分词、u 表示助词、ws 表示外文词、a 表示形容词。

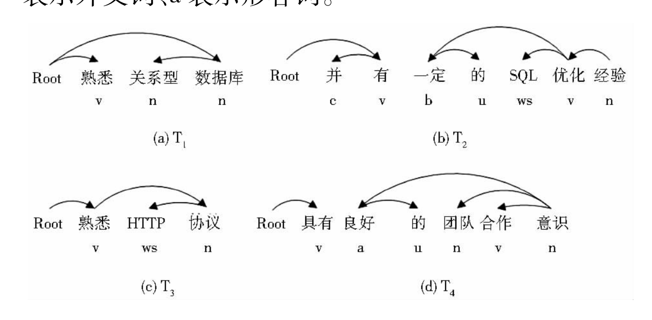


图 4　依存树示例

据此,本文将 Root 指向语句的核心动词剔除,保留剩余所有动词、名词、外文词等实词,采用 n-gram 策略,将频次大于 1 次,且长度在 1－4 的词串作为候选技能。表 1 为对图 4 利用依存句法分析选取的候选技能,并使用传统的方法进行比较,其中依存句法分析中 Root 指向的核心动词使用粗体表示。由表 1 可见,使用基于依存句法分析得到的候选技能包含更少的噪音词串,有效地过滤了"熟悉""具有""有"等词。

3.3 C-value 值计算

C-value 方法为每个候选技能计算术语性,统计信息包括候选技能的词频、词长、包含当前候选技能的更

表 1　候选技能选取方法比较示例

传统方法	基于依存句法分析的方法
熟悉	关系型
关系型	数据库
数据库	关系型 数据库
数据 关系型	SQL
关系型 数据库	优化
熟悉 关系型 数据库	经验
有	SQL 优化
SQL	优化 经验
优化	SQL 优化 经验
经验	HTTP
SQL 优化	协议
优化 经验	HTTP 协议
SQL 优化 经验	团队
HTTP	合作
协议	意识
熟悉 HTTP	团队 合作
HTTP 协议	合作 意识
熟悉 HTTP 协议	团队 合作 意识
具有	
团队	
合作	
意识	
团队 合作	
合作 意识	
团队 合作 意识	

长候选术语的频次和个数。C-value 值计算方法如公式(1)所示：

$$C\text{-}value(x)=\begin{cases}\log|x|\cdot tf_x^{(T)} & x\ \text{未被套}\\ \log|x|\cdot\left(tf_x^{(T)}-\frac{1}{|C_x|}\sum_{y\in CD_x}tf_y^{(T)}\right) & \text{其他}\end{cases}$$

公式(1)

其中，x 表示候选技能；$|x|$ 表示 x 的长度；$tf_x^{(T)}$ 表示 x 在网络招聘文本目标集 T 中出现的频次；C_x 表示网络招聘文本目标集包含 x 的候选技能集合；$|C_x|$ 表示集合 C_x 中元素个数。由公式(1)可知，C-value 与该候选技能在目标语料中频次有关，频次越高，其术语度越大。在此基础上，又考虑了候选技能的长度，认为长串出现频次比短串出现频次更有意义，是技能的可能性更大。

3.4　领域相关性度量

为了度量候选技能的领域相关性，本文首先衡量候选技能中每个词的领域相关性，然后依据每个词的领域相关性，计算得到候选技能词串的领域相关性。

3.4.1　词领域相关性度量　技能信息由若干词组成，因此，本文首先提出词领域相关性(domain relevance, DR)，描述词与特定领域的关联程度。具体地，给定目标领域网络招聘文本集 T 和包含非目标领域网络招聘文本集 NT，通过比较词 w 在目标领域网络招聘文本集 T 和非目标领域网络招聘文本集 NT 出现频次，其定义如公式(2)所示：

$$DR_w^{(T)}=\frac{p_w^{(T)}}{p_w^{(NT)}}\qquad \text{公式(2)}$$

其中，$p_w^{(T)}=\frac{tf_w^{(T)}}{|T|}$表示词 w 在目标领域网络招聘文本集 T 中出现的概率，$tf_w^{(T)}$ 表示词 w 在 T 中出现的频次，|T|表示目标领域网络招聘文本集包含的词数；类似地，$p_w^{(NT)}=\frac{tf_w^{(NT)}}{|NT|}$表示词 w 在非目标领域网络招聘文本集 NT 中出现的概率，$tf_w^{(NT)}$ 表示词 w 在 NT 中出现的频次，|NT|表示非目标领域网络招聘文本集包含单词数。由公式(2)可知，当 DR 值越大，表明词与目标领域越相关；反之，DR 值越小，表明词与目标领域越不相关。

3.4.2　候选技能领域相关性度量　候选技能包含若干个词 $x=\{w_1,w_2,\cdots,w_m\}$，根据词的领域相关性，度量候选技能词串与特定领域的相关程度。具体地，计算方法如公式(3)所示：

$$DR_x^{(T)}=\prod_{i=1}^{m}DR_{w_i}^{(T)}\qquad \text{公式(3)}$$

其中，$DR_x^{(T)}$ 表示基于候选技能 x 在目标领域 T 的领域相关程度；由定义可知，只有当候选技能中的每个词都具有较高领域相关性时，候选技能才具有较高的领域相关性。

3.5　融合领域相关性的 C-value 值计算

当候选技能 C-value 值越大，领域相关性越大，则越可能是技能。因此，本文提出融合领域相关性的 C-value 值计算，以度量候选术语成为技能的可能性。计算方法如公式(4)所示：

$$DRC\text{-}value(x)=DR_x^{(T)}\times C\text{-}value(x)\qquad \text{公式(4)}$$

最后，将融合领域相关性的 C-value 值进行降序排列，前若干个候选技能作为被抽取的技能。

4　实验

4.1　数据集

为了验证本文提出方法的可行性与有效性，实验抓取国内主流招聘网站“前程无忧”(www.51job.com)数据，抽取计算机领域的招聘文本技能信息。前程无忧是一家网络招聘服务提供商，是中国最具影响力的

人才招聘网站之一。按照职能,在前程无忧网站选取"计算机/互联网/通信/电子"职能抓取数据作为网络招聘文本目标集,数据抓取日期为 2018 年 3 月 19 日至 2018 年 3 月 26 日。在前程无忧招聘网站依次选取其他非计算机领域相关职能抓取数据,数据抓取日期为 2018 年 3 月 19 日至 2018 年 3 月 26 日。抓取后的网页文本去除本科以下学历、内容重复、全英文、没有写明任职要求的网络招聘文本,最后得到的数据基本信息如表 2 所示。

表 2　数据集基本信息

数据集类型	领域	网络招聘文本数
目标领域网络招聘文本集	计算机/互联网/通信/电子	10 000
非目标领域网络招聘文本集	保险	2 000
	会计、审计	2 000
	房地产	2 000
	建筑、建材、工程	2 000
	广告	2 000
	电气、电力、水利	2 000
	电子技术、半导体、集成电路	2 000
	服装、纺织、皮革	2 000
	机械、设备、重工	2 000
	家居、室内设计、装潢	2 000
	家居、家电、玩具、礼品	2 000
	检测、认证	2 000
	教育、培训	2 000
	金融、投资、证券	2 000
	酒店、旅游	2 000

4.2　实验步骤与评估标准

实验首先对目标领域网络招聘文本集和非目标领域网络招聘文本集进行文本预处理。利用依存句法分析,选取候选技能。利用非目标领域网络招聘文本集,计算候选技能中每个词的领域相关性,以获得候选术语的领域相关性。最终将其融入候选技能 C-value 值之中,按值降序排列,选取前 N 个候选技能作为被抽取的技能。

人工判定前 N 个候选技能信息是否正确,从而计算出方法的准确率。同时,从目标领域网络招聘文本集中随机抽取 500 篇招聘文本,人工识别其中的技能信息,检验方法的召回率。最后,结合准确率和召回率得到 F 值指标,以评估方法。准确率、召回率和 F 值计算方法见公式(5)-公式(7):

$$准确率 = \frac{正确抽取的技能信息数}{抽取出的技能信息数} \times 100\% \qquad 公式(5)$$

$$召回率 = \frac{正确抽取的技能信息数}{正确技能信息数} \times 100\% \qquad 公式(6)$$

$$F值 = 2 \times \frac{准确率 \times 召回率}{准确率 + 召回率} \times 100\% \qquad 公式(7)$$

4.3　实验结果

4.3.1　基于依存句法分析的候选技能选取的有效性　实验首先验证基于依存句法分析的候选技能选取的有效性。为此,实验分别使用传统的候选技能选取方法和本文的基于依存句法分析的候选技能选取方法,然后均使用 C-value 值排序候选技能。前一种方法称为 C-value,后者称为 DepC-value。实验比较结果见图 5-图 7。

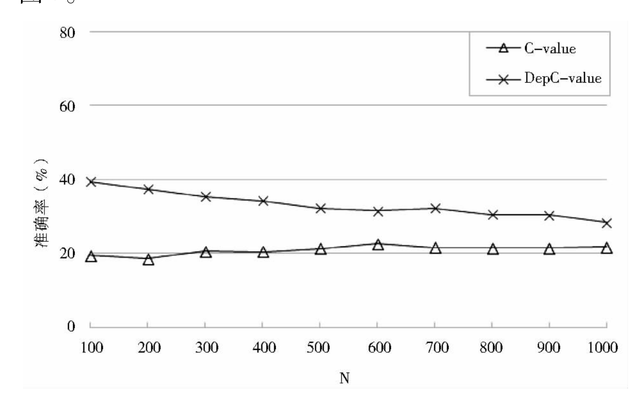


图 5　C-value 与 DepC-value 方法准确率比较

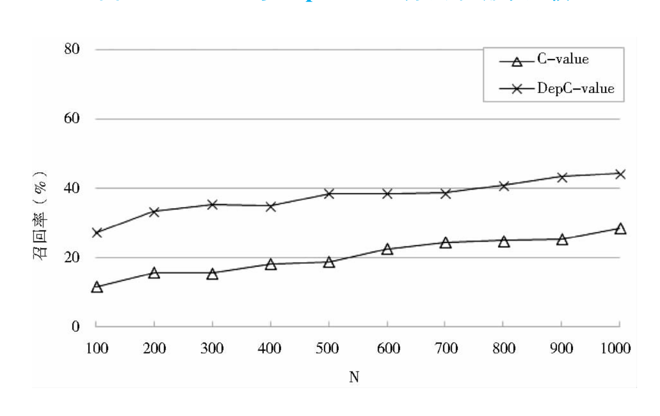


图 6　C-value 与 DepC-value 方法召回率比较

由图 5-图 7 可见,DepC-value 方法准确率、召回率和 F 值高于 C-value 方法,这表明使用基于依存句法分析选取候选术语的有效性。基于依存句法分析的候选技能选取方法针对网络招聘文本技能要求语句通常为动宾语句的特点,过滤掉不必要的噪声动词,从而提高了技能抽取的准确率和召回率。特别地,该方法还大幅度减少了候选技能的数目,从而缩短了后续计算时间。

表 3 为频次最高的前 10 个通过依存句法分析被剔除的噪声动词。由表 3 可见,通过依存句法分析,可

图书情报工作
第 63 卷 第 13 期 2019 年 7 月

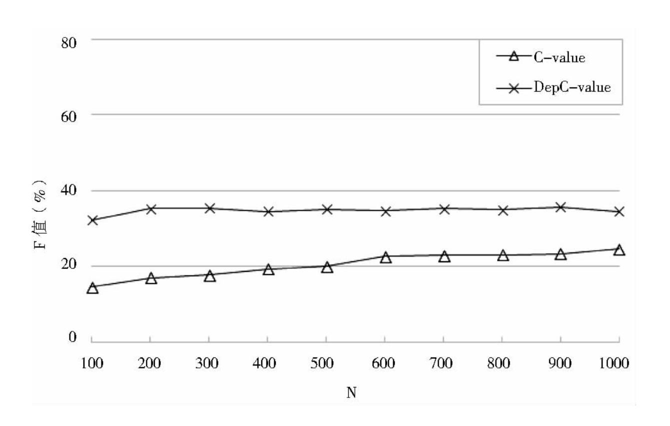


图 7 C-value 与 DepC-value 方法 F 值比较

以准确过滤噪声动词，从而减少不正确的候选技能。表 4 为 C-value 方法和 DepC-value 方法的前 10 个被抽取技能。其中，基于依存句法分析剔除的动词使用粗体表示。由表 4 可以看到，由于传统的候选技能术语无法过滤不必要的动词，而含有这些动词词串的候选技能大量存在，不仅造成计算时间长，也造成 C-value 值高，从而降低了技能抽取的准确率和召回率。相反，使用基于依存句法分析的候选术语选取方法，则能有效地过滤一部分噪声动词，从而获得更高的技能抽取准确率和召回率。

表 3 前 10 个被剔除的噪声动词

序号	被剔除的噪声动词	例句
1	有	有 arcgis 开发经验
2	熟悉	熟悉 web 接口
3	具有	具有 1 – 3 年的数据仓库测试经验
4	具备	具备服务器的部署配置能力
5	了解	了解 js 模块化
6	理解	深入理解软件架构及设计模式
7	掌握	熟练掌握网络通信
8	对	对分布式储存计算有较深入了解
9	包括	包括前端技术
10	拥有	拥有 erp 项目经验

表 4 C-value 和 DepC-value 方法的前 10 被抽取技能

序号	C-value	DepC-value
1	相关 专业	相关 专业
2	工作 经验	工作 经验
3	有 良好	团队 合作 精神
4	熟练 使用	沟通 能力
5	团队 合作 精神	学习 能力
6	沟通 能力	需求 分析
7	具有 良好	Java 开发
8	有 较强	相关 工作 经验
9	有 一定	责任心 强
10	具备 良好	团队 协作 能力

4.3.2 领域相关性度量的有效性 接着，实验评估了领域相关性度量的有效性。在上一个实验 DepC-value 方法的基础上，使用领域相关性结合 C-value 方法重新评估候选技能，得到 DepDRC-value 方法。

DepC-value 方法与 DepDRC-value 方法的比较结果见图 8 – 图 10。由图 8 – 图 10 可见，DepDRC-value 方法准确率、召回率和 F 值明显高于 DepC-value 方法，表明融入候选技能的领域相关性度量，能够明显提高技能抽取的准确率、召回率和 F 值。

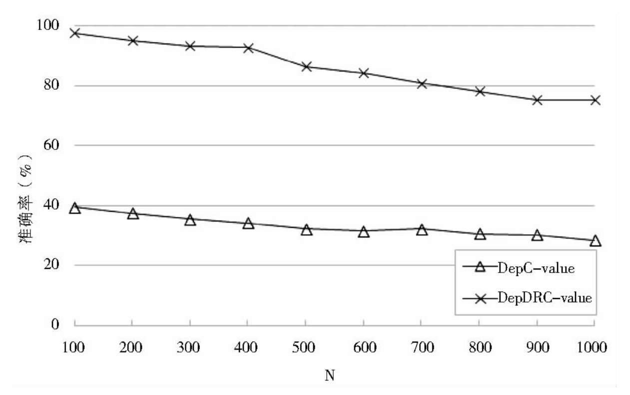


图 8 DepC-value 与 DepDRC-value 方法准确率比较

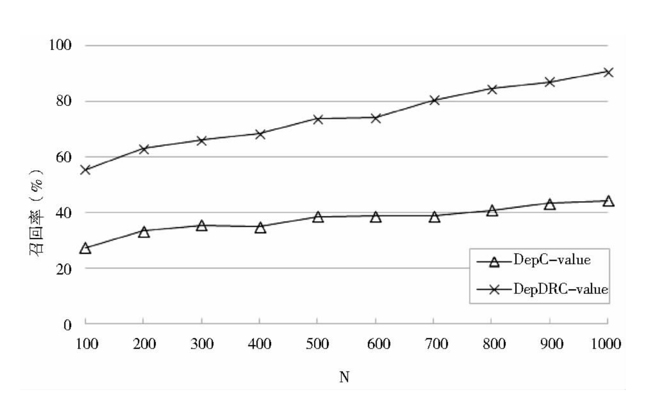


图 9 DepC-value 与 DepDRC-value 方法召回率比较

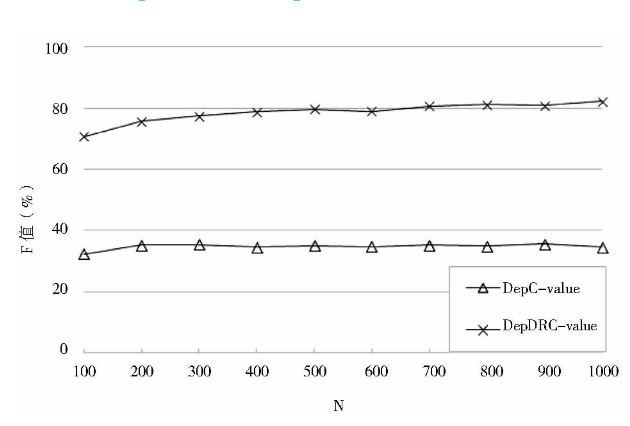


图 10 DepC-value 与 DepDRC-value 方法 F 值比较

表 5 显示 DepC-value 方法和 DepDRC-value 方法的前 10 个被抽取技能。其中，正确的技能使用粗体表示。由表 5 可见，DepDRC-value 方法通过度量候选技

能的领域相关性,从而有效降低低领域相关候选术语 DRC-value 值,如“相关 专业”“工作 经验”等。而一些具有较高领域相关度的候选技能,如“SQL 语句”“Linux 常用命令”等,增加了其 DRC-value 值,从而使 DepDRC-value 方法获得更高的技能抽取准确率、召回率和 F 值。

表 5　DepC-value 和 DepDRC-value 方法的前 10 被抽取技能

序号	DepC-value	DepDRC-value
1	相关 专业	**SQL 语句**
2	工作 经验	**Linux 常用命令**
3	团队 合作 精神	**关系数据库 MySQL**
4	沟通 能力	**SQL 查询语句**
5	学习 能力	**MySQL 主从复制**
6	需求 分析	**SQL 关系数据库**
7	**Java 开发**	**关系数据库 SQL**
8	相关 工作 经验	**JavaScript 程序模块**
9	责任心 强	**MySQL 关系数据库**
10	团队 协作 能力	**Lamada 表达式**

4.3.3　与其他方法比较　为验证本文提出方法的有效性,比较以下 4 种方法:①C-value:使用传统方法选取候选技能,根据 C-value 值度量候选技能;②MIC-value:使用传统方法选取候选技能,将候选技能的互信息融入 C-value 之中,形成 MIC-value 方法[20];③EnC-value:使用传统方法选取候选技能,将候选技能词串的邻接熵融入 C-value 值之中,形成 EnC-value 方法[20];④DepDRC-value:本文提出的方法,首先使用基于依存句法分析选取候选技能,然后融入候选技能的领域相关性信息,形成 DepDRC-value 方法。

实验结果如图 11－图 13 所示。由图 11－图 13 可见,DepDRC-value 方法的准确率、召回率和 F 值均明显高于其他几种方法,这表明 C-value、MIC-value、EnC-value 方法并不适合于网络招聘文本技能抽取。本文提出的 DepDRC-value 方法针对招聘网络文本的特点,利用依存句法分析,融合领域相关性信息能够大幅度提高 C-value 的准确率、召回率和 F 值。

表 6 列出了 MIC-value、EnC-value 和 DepDRC-value3 种方法排序的前 10 个被抽取技能,正确技能使用粗体表示。由表 6 可见,MIC-value 方法使用候选技能词串的互信息衡量候选技能中词的紧密程度。但是,由于一些非技能词串也高频出现,使得这些词串的互信息值较高,从而导致错误的抽取结果。EnC-value 方法使用邻接熵衡量候选术语左右邻接词的不确定性,其不确定性越大,表明其邻接词包含的信息越

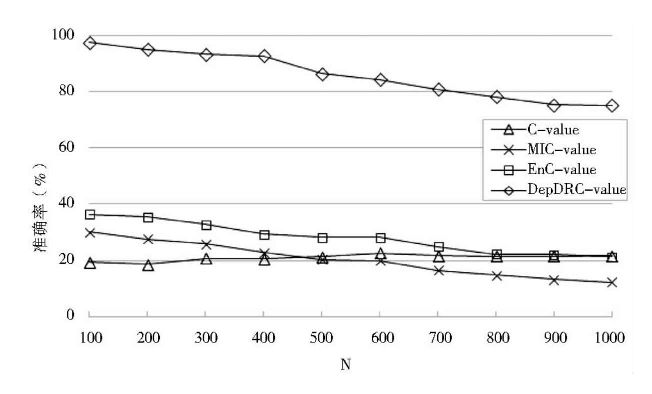


图 11　4 种方法准确率比较

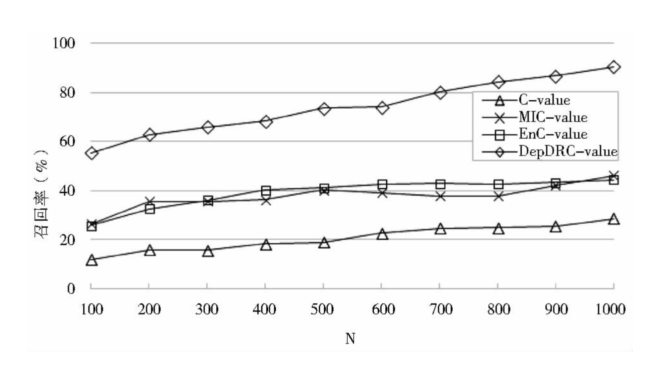


图 12　4 种方法召回率比较

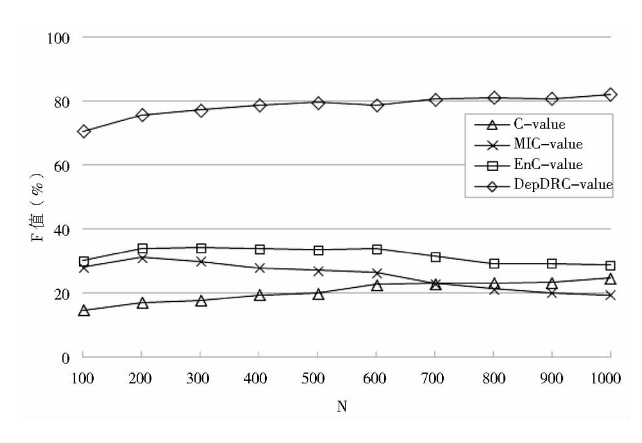


图 13　4 种方法 F 值比较

多,越可能是术语。然而一些非技能候选词串其邻接信息丰富,如“熟练 使用”可以连接很多种信息,具有较高的邻接熵,造成错误的抽取结果。DepDRC-value 利用依存句法分析,通过引入网络招聘文本辅助集,能够很好地度量选技能的领域相关性,克服 C-value 方法的缺点,从而提高技能抽取的准确率和召回率。

5　结论

网络招聘信息中常含有企业对所招岗位技能需求的具体描述,反映了当前就业市场对人才的技能需求。因此,通过分析网络招聘信息,可以了解整个社会对某

表 6 3 种方法前 10 个被抽取技能

序号	MIC-value	EnC-value	DepDRC-value
1	相关 专业	熟练 使用	**SQL 语句**
2	工作 经验	团队 合作	**Linux 常用命令**
3	团队 合作 精神	具有 良好	**关系数据库 MySQL**
4	沟通 能力	以上 工作 经验	**SQL 查询语句**
5	学习 能力	独立 完成	**MySQL 主从复制**
6	需求 分析	能 独立	**SQL 关系数据库**
7	**Java 开发**	熟悉 常用	**关系数据库 SQL**
8	相关 工作 经验	**SQL 语句**	**JavaScript 程序模块**
9	责任心 强	能够 独立	**MySQL 关系数据库**
10	团队 协作 能力	**SQL 数据库**	**Lamada 表达式**

领域人才的技能需求。然而,网络招聘信息往往为非结构化文本,传统的技能需求分析方法通常需要手工抽取招聘文本中的技能,以进行技能需求分析。显然,手工抽取招聘文本中的技能信息很难满足大数据量、非结构化环境下的网络招聘信息分析要求。本文针对网络招聘文本的特点,利用依存句法分析选取候选技能,然后提出技能的领域相关性概念,将候选技能领域相关性融入 C-value 方法之中,以自动抽取网络招聘文本中的技能。实验表明,本文提出的方法针对大数据网络招聘信息,能够从海量网络招聘文本中自动、快速、准确地抽取技能。进一步地,未来工作将尝试依据网络招聘文本中抽取的技能信息,进行热门招聘岗位技能需求分析,为学生、教师、高校提供有指导性的岗位技能需求信息。

参考文献:

[1] WOWCZKO I. Skills and vacancy analysis with data mining techniques[J]. Informatics, 2015, 2(4):31 – 49.

[2] KIM J Y, LEE C K. An empirical analysis of requirements for data scientists using online job postings [J]. International journal of software engineering and its application, 2016, 10(4): 161 – 172.

[3] MAURO A D, GRECO M, GRIMALDI M, et al. Beyond data scientists: a review of big data skills and job families[C]//Proceedings of the 2016 international forum on knowledge asset dynamics. Berlin: Springer International Publishing, 2016: 1844 – 1857.

[4] 吕斌,张通,周珏. 面向组织的具有通用性的情报职业及情报从业人员——基于组织招聘网页信息挖掘的分析之一[J]. 图书情报工作, 2009, 53(4): 19 – 23.

[5] 李国秋,桑培铭. 情报过程——情报职业的核心:问题域及方法论——基于组织招聘网页信息挖掘的分析之二[J]. 图书情报工作, 2009, 53(4): 24 – 27.

[6] 夏火松, 潘筱听. 基于 Python 挖掘的大数据学术研究与人才需求的关系研究[J]. 信息资源管理学报, 2017, 7(1): 4 – 12.

[7] 黄崑, 王凯飞, 王珊珊,等. 数据类岗位招聘需求调查及对图情学科人才培养的启示[J]. 图书情报知识, 2016,6(1):42 – 53.

[8] FRANTZI K, ANANIADOU S, MIMA H. Automatic recognition of multi-word terms:. the C-value/NC-value, method[J]. International journal on digital libraries, 2000, 3(2):115 – 130.

[9] SODHI M S, SON B G. Content analysis of OR job advertisements to infer required skills[J]. The journal of the Operational Research Society, 2010, 9(1): 1315 – 1327.

[10] ZHAO M, JAVED F, JACOB F, et al. l SKILL: a system for skill identification and normalization[C]// Proceedings of the twenty-seventh conference on innovative applications of artificial intelligence. Palo Alto: AAAI, 2015: 4012 – 4017.

[11] XU T, ZHU H, ZHU C, et al. Measuring the popularity of job skills in recruitment market: a multi-criteria approach[C]//Proceedings of the 32nd AAAI conference on artificial intelligence. Menlo Park: AAAI, 2018: 3013 – 3028.

[12] 詹川. 基于文本挖掘的专业人才技能需求分析——以电子商务专业为例[J]. 图书馆论坛, 2017, 5(1): 116 – 123.

[13] 夏立新, 楚林, 王忠义,等. 基于网络文本挖掘的就业知识需求关系构建[J]. 图书情报知识, 2016, 169(1):94 – 100.

[14] BASTIAN M, HAYES M, VAUGHAN W, et al. LinkedIn skills: large-scale topic extraction and inference[C]// ACM conference on recommender systems. New York: ACM, 2014:1 – 8.

[15] 刘睿伦, 叶文豪, 高瑞卿,等. 基于大数据岗位需求的文本聚类研究[J]. 数据分析与知识发现, 2017, 12(12): 32 – 40.

[16] CONRADO M D, PARDO T A, REZENDE S O. A machine learning approach to automatic term extraction using a rich feature set [C]// The North American chapter of the Association for Computational Linguistics. Stoudsburg: ACL, 2013: 16 – 23.

[17] PIAO S, FORTH J, GACITUA R, et al. Evaluating tools for automatic concept extraction: a case study from the musicology domain [C]//Proceedings of digital features. Piscataway: IEEE, 2010: 78 – 85.

[18] SPASIC I, GREENWOOD M, PREECE A, et al. FlexiTerm: a flexible term recognition method[J]. Journal of biomedical semantics, 2013, 4(1):27 – 42.

[19] MAYNARD D, ANANIADOU S. Identifying terms by their family and friends[C]// Proceeding of the 18th conference on computational linguistics. New York: ACM, 2000:530 – 536.

[20] 周霜霜, 徐金安, 陈钰枫,等. 融合规则与统计的微博新词发现方法[J]. 计算机应用, 2017, 37(4):1044 – 1050.

[21] 赵京胜, 朱巧明, 周国栋,等. 自动关键词抽取研究综述[J]. 软件学报, 2017, 28(9):2431 – 2449.

[22] 刘怀军,车万翔,刘挺. 中文语义角色标注的特征工程[J]. 中文信息学报, 2007, 21(1): 79 – 84.

[23] 哈尔滨工业大学语言技术平台 LTP[EB/OL]. [2018 – 12 – 30]. http://ir. hit. edu. cn/demo/ltp.

[24] CHE W, LI Z, LIU T, A Chinese language technology platform

[C]//The 23th international conference on computational linguistics. Stroudsburg: ACL, 2010: 3－16.

作者贡献说明:

俞琰:提出研究思路,设计研究方案,进行试验,撰写论文;

陈磊:数据清洗;

姜金德:分析数据,修改论文;

赵乃瑄:修改论文。

Research on Skill Information Automatic Extraction from Online Recruitment Texts

Yu Yan[1,2] Chen Lei[1] Jiang Jinde[3] Zhao Naixuan[1]

[1] Information Service Department, Nanjing Tech University, Nanjing 210009

[2] Computer Science Department, Chengxian College, Southeast University, Nanjing 211816

[3] School of Business, Nanjing Xiaozhuang University, Nanjing 211171

Abstract: [**Purpose/significance**] Aiming at the problem that the current manual skill information extraction from the online recruitment post is not suitable for the analysis of large data volume information, this paper proposes an automatic skill information extraction for a large number of online recruitment texts. [**Method/process**] According to the characteristics of online recruitment texts, the candidate skills are analyzed by dependency syntax analysis, then the domain relevance indicators are used to measure candidate skills, and they are integrated into the traditional terminology extraction method to form a method for automatic extraction of skill information from online recruitment texts. [**Result/conclusion**] Experiments show that the proposed method can extract skill information automatically, quickly and accurately from the mass online recruitment texts.

Keywords: online recruitment text　skill information automatic extraction　term extraction